

CONSTANT TIME EXPECTED SIMILARITY ESTIMATION USING STOCHASTIC OPTIMIZATION

MARKUS SCHNEIDER^{*†}, WOLFGANG ERTEL[†] & GÜNTHER PALM^{*}

Abstract

A new algorithm named *EX*pected *Si*milarity *E*stimation (EXPoSE) was recently proposed to solve the problem of *large-scale anomaly detection*. It is a non-parametric and distribution free kernel method based on the Hilbert space embedding of probability measures. Given a dataset of n samples, EXPoSE needs only $\mathcal{O}(n)$ (linear time) to build a model and $\mathcal{O}(1)$ (constant time) to make a prediction. In this work we improve the *linear* computational complexity and show that an ϵ -accurate model can be estimated in *constant* time, which has significant implications for large-scale learning problems. To achieve this goal, we cast the original EXPoSE formulation into a stochastic optimization problem. It is crucial that this approach allows us to determine the number of iteration based on a desired accuracy ϵ , *independent of the dataset size* n . We will show that the proposed stochastic gradient descent algorithm works in general (possible infinite-dimensional) Hilbert spaces, is easy to implement and requires no additional step-size parameters.

1 Introduction

*EX*pected *Si*milarity *E*stimation (EXPoSE) was recently proposed to solve the problem of large-scale anomaly detection, where the number of training samples n and the dimension of the data d are too high for most other algorithms [SEP15]. Here, “anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as *anomalies*” [CBK09].

As explained later in detail, the EXPoSE anomaly detection classifier

$$\eta(y) = \langle \phi(y), \mu[\mathbb{P}] \rangle$$

calculates a score (the likelihood of y belonging to the class of normal data) using the inner product between a feature map ϕ and the kernel mean map $\mu[\mathbb{P}]$ of the distribution \mathbb{P} (Fig. 1). Given a training dataset of size n , the authors provide a methodology to train this classifier in $\mathcal{O}(n)$ time and show that calculating a score for a query point can be done in $\mathcal{O}(1)$ time. The question arises if it is possible to improve on the linear training time and create an algorithm which is

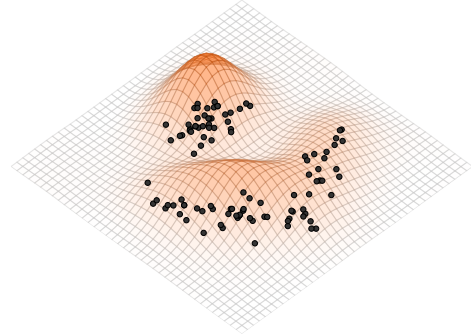


Figure 1: Sketch of the EXPoSE scores $\eta(y)$ in \mathbb{R}^2 , given some samples (black dots).

completely independent of the dataset size.

The answer to this question is positive if a high accuracy sample estimate of $\mu[\mathbb{P}]$ does not improve the anomaly detection performance. As Bousquet and Bottou [BB08] observed, for most machine learning applications there is no need to optimize below the statistical error. The authors argue that accurately minimizing an empirical cost function does not gain much since it is itself an approximation of the expected costs and therefore contains errors. We will see that it is possible to determine the number of samples needed to achieve a desired accuracy (the maximal deviation from the optimal model) of EXPoSE without any dependence on the datasets size n .

^{*} Institute of Neural Information Processing,
University of Ulm, Germany

[†] Institute for Artificial Intelligence,
Ravensburg-Weingarten University of Applied Sciences, Germany

1.1 CONTRIBUTIONS & RELATED WORK

In this work we derive a methodology to build an ϵ -accurate model w of $\mu[\mathbb{P}]$ using only a random subset of the training data by means of stochastic optimization.

Definition 1: We say an algorithm finds an ϵ -accurate solution w of an objective function f if

$$f(w) \leq \inf f + \epsilon$$

for a given $\epsilon > 0$. «

We will show that for the proposed objective function $\mathbb{E}[f(w_t) - f(\mu[\mathbb{P}])] \leq \mathcal{O}(1/t)$, where w_t only needs access to t random dataset elements, $t \in \{1, 2, \dots, n\}$. The key observation is that for a given $\epsilon > 0$ we can reach $\|w_t - \mu[\mathbb{P}]\| < \epsilon$ in a *fixed* number of iterations *independent* of the dataset size. Moreover, it can be shown that (without further assumptions) the $\mathcal{O}(1/t)$ rate is optimal for stochastic optimization [Aga+11].

Due to the low iteration costs, stochastic optimization and especially stochastic gradient (SG) methods [BB08; RSB12], are widely used for training machine learning models on very large-scale datasets. Such algorithms are used for example to train support vector machines [SS+11], logistic regression [Bac14] and lasso models [SST11]. However, *this is the first time that EXPoSE is considered as an optimization problem* and we will show that the derived algorithms is of general interest for applications of the kernel mean map $\mu[\mathbb{P}]$.

Other optimization techniques such as projected gradient decent [BV04] or Nesterov’s accelerated gradient descent [Nes83; Neso04] are also applicable in principle, however a single gradient evaluation takes already $\mathcal{O}(n)$ time and hence would be slower than the originally proposed EXPoSE approach. Other stochastic gradient methods [RSB12] can obtain a better convergence rate than $\mathcal{O}(1/t)$ for an objective composed of a sum of smooth functions. However this requires multiple passes over the datasets is therefor of no benefit.

2 Problem Description

EXPoSE is a probabilistic approach which assumes that the *normal*, non-anomalous data is distributed according to some measure \mathbb{P} . More formally, let X be a random variable taking values in a measure space $(\mathcal{X}, \mathcal{H})$ with distribution \mathbb{P} . We denote the reproducing kernel Hilbert space (RKHS) associated with the kernel $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ with $(\mathcal{H}, \langle \cdot, \cdot \rangle)$. A RKHS is a Hilbert space of functions $g: \mathcal{X} \rightarrow \mathbb{R}$, where the evaluation functional $\delta_x: g \mapsto g(x)$ is continuous. The function $\phi: \mathcal{X} \rightarrow \mathcal{H}$

with

$$k(x, y) = \langle \phi(x), \phi(y) \rangle$$

is called *feature map* denoted by $\phi(x) = k(x, \cdot)$. Throughout the paper, we use $\|\cdot\|_{\mathcal{H}}$ to denote the norm induced by the inner product defined as $\|g\|_{\mathcal{H}} = \sqrt{\langle g, g \rangle}$.

EXPoSE calculates a score which can be interpreted as the likelihood of a query point belonging to the distribution of normal data \mathbb{P} . This is done in the following way.

Definition 2: The *expected similarity* of $y \in \mathcal{X}$ to the (probability) distribution \mathbb{P} is defined as

$$\eta(y) = \int_{\mathcal{X}} k(y, x) d\mathbb{P}(x),$$

where $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a reproducing kernel. «

Intuitively speaking the query point y is compared to all other points of the distribution \mathbb{P} . It can be shown [SEP15] that this equation can be rewritten as an inner product between the feature map $\phi(y)$ and the kernel embedding $\mu[\mathbb{P}]$ of \mathbb{P} as

$$\begin{aligned} \eta(y) &= \int_{\mathcal{X}} k(y, x) d\mathbb{P}(x) \\ &= \langle \phi(y), \mu[\mathbb{P}] \rangle, \end{aligned}$$

where the kernel embedding is defined as follows.

Definition 3: The kernel embedding or kernel mean map $\mu[\mathbb{P}]$ associated with the continuous, bounded and positive-definite kernel function k is

$$\mu[\mathbb{P}] = \int_{\mathcal{X}} \phi(x) d\mathbb{P}(x),$$

where \mathbb{P} is a Borel probability measure on \mathcal{X} . «

To facilitate the further analysis, we assume that the kernel k is measurable and bounded such that $\mu[\mathbb{P}]$ exists for all $\mathbb{P} \in \mathcal{M}_+^1(\mathcal{X})$ [SFL11]. Since the underlying distribution \mathbb{P} is in general unknown and only a set of $n \in \mathbb{N}$ samples $\{x_1, \dots, x_n\}$ from \mathbb{P} is available for analysis, the empirical measure

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i},$$

act as a surrogate, where δ_x is the Dirac measure. \mathbb{P}_n can be used to construct an approximation $\mu[\mathbb{P}_n]$ of $\mu[\mathbb{P}]$ as

$$\mu[\mathbb{P}] \approx \mu[\mathbb{P}_n] = \int_{\mathcal{X}} \phi(x) d\mathbb{P}_n(x) = \frac{1}{n} \sum_{i=1}^n \phi(x_i)$$

which is called *empirical kernel embedding* [Smo+07].

The consequence of the equation above is, that the empirical kernel embedding $\mu[\mathbb{P}_n]$ has a computational complexity with linear dependence on n and responsible for the *linear* EXPoSE training time. Next, we will look at the EXPoSE classifier from the perspective of a stochastic optimization problem to deliver an ϵ -accurate approximation of $\mu[\mathbb{P}]$ in *constant* time. A reduction of the computationally complexity from linear to constant for the empirical kernel mean map has significant impact on a variety of applications based on the kernel embedding such as for example statistical hypotheses testing [Gre+12] or independence testing [Gre+05].

However the main focus of this work is to improve the EXPoSE training time from linear to constant.

3 Stochastic Optimization

This sections derives the stochastic optimization problem together with some general conditions which will be necessary at a later stage. Obviously $\mu[\mathbb{P}] \in \mathcal{H}$ is the solution of the following unconstrained optimization problem

$$\begin{aligned} \min_{w \in \mathcal{H}} g(w) &= \min_{w \in \mathcal{H}} \|\mu[\mathbb{P}] - w\|_{\mathcal{H}}^2 \\ &= \min_{w \in \mathcal{H}} \langle w, w \rangle - 2\langle \mu[\mathbb{P}], w \rangle + \langle \mu[\mathbb{P}], \mu[\mathbb{P}] \rangle \\ &= \min_{w \in \mathcal{H}} \frac{1}{2} \langle w, w \rangle - \langle \mu[\mathbb{P}], w \rangle. \end{aligned}$$

This is equivalent to the *stochastic optimization problem*, where we minimize over the expectation of an objective function

$$\min_{w \in \mathcal{H}} \mathbb{E}[f(w)] = \min_{w \in \mathcal{H}} \int_{\mathcal{X}} f(w) \, d\mathbb{P}(x),$$

with

$$f(w) = \frac{1}{2} \langle w, w \rangle - \langle \phi(X), w \rangle,$$

where the expectation is taken with respect to the random variable X .

We will assume that we can generate independent samples from \mathbb{P} and furthermore require an oracle which returns a *stochastic subgradient* $\tilde{\nabla}f(w)$ of f at w . A stochastic subgradient has the property that

$$\mathbb{E}[\tilde{\nabla}f(w)] = \nabla f(w) \in \partial f(w)$$

which means its expectation is equal to a subgradient $\nabla f(w)$. Here $\partial f(w)$ denotes the set of all subgradients at w called the *subdifferential* which is a subset of the dual

\mathcal{H}^* of \mathcal{H} defined by

$$\partial f(x) = \{\xi^* \in \mathcal{H}^* \mid f(y) - f(x) \geq \xi^*(y - x)\}.$$

Proposition 1: The random variable

$$\tilde{\nabla}f(w) = w - \phi(X)$$

is a stochastic unbiased gradient of f at w . «

Proof: The expectation of $\tilde{\nabla}f(w)$ is given by

$$\begin{aligned} \mathbb{E}[\tilde{\nabla}f(w)] &= \int w - \phi(X) \, d\mathbb{P}(x) \\ &= w - \mu[\mathbb{P}] \in \partial f(w) \end{aligned}$$

which is a stochastic unbiased (sub)gradient by definition. □

We are going to solve this optimization problem with the *stochastic approximation* algorithm [RM51] described next.

3.1 STOCHASTIC APPROXIMATION

Let \mathcal{H} be a Hilbert space, $H \subseteq \mathcal{H}$ be a subset and $f: H \rightarrow \mathbb{R}$ some objective function. Furthermore let

$$\Pi_H(w) = \arg \min_{v \in H} \|w - v\|_{\mathcal{H}}$$

be the metric projection operator. Π_H is in general nonexpanding such that

$$\|\Pi_H(w) - \Pi_H(w')\|_{\mathcal{H}} \leq \|w - w'\|_{\mathcal{H}}$$

holds. Then the classic stochastic approximation algorithm [RM51] creates the sequence (w_t) as

$$w_{t+1} = \Pi_H(w_t - \gamma_t \tilde{\nabla}f(w_t)),$$

to solve the stochastic optimization problem

$$\min_{w \in H} \mathbb{E}[f(w)]$$

starting at some $w_1 \in H$. Here (γ_t) is a sequence of positive step sizes and the optimal solution to the problem is denoted by w^* .

Nemirovski et al. [Nem+09] considered $\mathcal{H} = \mathbb{R}^d$ and showed that stochastic approximation can obtain a $\mathcal{O}(1/t)$ convergence rate if the objective function f is differentiable and α -strongly convex on H . Here, α -strongly convex means there exists a constant $\alpha > 0$ such that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \alpha \|y - x\|_{\mathcal{H}}^2$$

for all $x, y \in H$. An additional requirement is that the stochastic subgradient has to be bounded in expectation

$$\mathbb{E}[\|\tilde{\nabla}f(w)\|_{\mathcal{H}}^2] \leq M^2 \quad \forall w \in H, M > 0$$

and the step sizes need to be $\gamma_t = \frac{\theta}{t}$ for some $\theta > \frac{1}{2\alpha}$. Under these conditions, Nemirovski et al. demonstrated that

$$\mathbb{E}[\|w_t - w^*\|_{\mathcal{H}}^2] \leq \frac{Q(\theta)}{t} \quad (1)$$

where

$$Q(\theta) = \max\left\{\theta^2 M^2 (2\alpha\theta - 1)^{-1}, \|w_1 - w^*\|_{\mathcal{H}}^2\right\}.$$

Furthermore, if the gradient is Lipschitz continuous, i.e. there is a constant $\beta > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\|_{\mathcal{H}} \leq \beta \|x - y\|_{\mathcal{H}}$$

for all $x, y \in H$, then

$$\mathbb{E}[f(w_t) - f(w^*)] \leq \frac{1}{2} \frac{\beta Q(\theta)}{t}. \quad (2)$$

For Lipschitz continuous strongly convex functions, the $\mathcal{O}(1/t)$ rate of convergence is unimprovable [Aga+11].

We will see that the bound from Nemirovski et al. does also hold when \mathcal{H} is a (possibly infinite-dimensional) Hilbert space as in the problem considered in this work. However, some care has to be taken since, unlike in finite-dimensional spaces, being closed and bounded does not imply that a set is compact when \mathcal{H} is infinite-dimensional. We also refer to [JN10] for a discussion on primal-dual subgradient methods in non-Euclidean spaces.

4 Stochastic Optimization of EXPoSE

In this section we show the existence and uniqueness of a solution for the previously defined stochastic optimization problem of EXPoSE and also that it meets all requirements for a $\mathcal{O}(1/t)$ convergence rate.

In the following let \mathcal{H} be a RKHS space with a bounded kernel k such that $\|k(x, y)\| \leq M^2$. Let $H \subseteq \mathcal{H}$ be a weakly sequentially closed and bounded set with $\|H\|_{\mathcal{H}} \leq M$. It is not hard to show the existence of a minimizer of

$$\min_{w \in H} \mathbb{E}[f(w)] = \min_{w \in H} \int \frac{1}{2} \langle w, w \rangle - \langle \phi(X), w \rangle d\mathbb{P}, \quad (3)$$

since we already know the solution $w^* = \mu[\mathbb{P}]$ assuming

that $\mu[\mathbb{P}] \in H$. This assumption holds since

$$\begin{aligned} \|\mu[\mathbb{P}]\|_{\mathcal{H}}^2 &= \left\| \int_{\mathcal{X}} \phi(x) d\mathbb{P}(x) \right\|_{\mathcal{H}}^2 \\ &\leq \int_{\mathcal{X}} \|\phi(x)\|_{\mathcal{H}}^2 d\mathbb{P}(x) \\ &= \int_{\mathcal{X}} k(x, x) d\mathbb{P}(x) \\ &\leq M^2. \end{aligned}$$

This solution is also unique. The proof requires $f(w)$ to be strongly convex, which is subject of the following property:

Proposition 2: The objective function $f(w)$ is α -strongly convex and its gradient is β -Lipschitz with $\alpha = \beta = 1$.

Proof: A function f is α -strongly convex if and only if $w \mapsto f(w) - \frac{\alpha}{2} \|w\|_{\mathcal{H}}^2$ is convex.

$$\begin{aligned} f(w) - \frac{1}{2} \|w\|_{\mathcal{H}}^2 &= \frac{1}{2} \langle w, w \rangle - \langle \phi(X), w \rangle - \frac{1}{2} \|w\|_{\mathcal{H}}^2 \\ &= -\langle \phi(X), w \rangle \end{aligned}$$

which is convex in w . Hence $\alpha = 1$.

Furthermore $Df(w): z \mapsto \langle w - \phi(X), z \rangle$ is the Fréchet derivative of f at w since

$$\lim_{h \rightarrow 0} \frac{\|f(w+h) - f(w) - \langle Df(w)|h \rangle\|}{\|h\|_{\mathcal{H}}} = 0$$

with dual pairing $\langle \cdot, \cdot \rangle$. The gradient $\nabla f(w) = w - \phi(X)$ is β -Lipschitz since

$$\begin{aligned} \|\nabla f(w) - \nabla f(v)\|_{\mathcal{H}} &= \|w - \phi(X) - v + \phi(X)\|_{\mathcal{H}} \\ &= \|w - v\|_{\mathcal{H}} \end{aligned}$$

for all $w, v \in H$ due to Riesz representation. \square

Besides the existence of a minimizer, its uniqueness plays an important role. The sufficient conditions for w^* to be unique are given by [Pey15, Corollary 2.19] which states the following:

Corollary 1: Let \mathcal{H} be reflexive. If $f: \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ is proper, convex, coercive and lower-semicontinuous, then $\arg \min f$ is nonempty and weakly compact. If, moreover f is strictly convex, then $\arg \min f$ is a singleton. \llcorner

Proof: w^* is unique: All Hilbert spaces are reflexive. Since f is continuous, proper ($\text{dom}(f) \neq \{\}$) and strongly convex it is also convex, coercive and lower-semicontinuous. \square

Next we state the two main theorems of this paper.

Theorem 1: Using the sequence

$$w_{t+1} = \Pi_H(w_t - \gamma_t \tilde{\nabla}f(w_t)),$$

with f given by Eq. (3) we have

$$\mathbb{E}[\|w_t - w^*\|_{\mathcal{H}}^2] \leq \frac{M^2}{t}$$

for all $t \in \mathbb{N}$. «

Proof: Since $Q(\theta)$ attains its optimal value at $\theta = 1/\alpha$ we get from Eq. (1) that

$$\mathbb{E}[\|w_t - w^*\|_{\mathcal{H}}^2] \leq t^{-1} \max\{\alpha^{-2}M^2, \|w_1 - w^*\|_{\mathcal{H}}^2\}$$

and we have

$$\mathbb{E}[\|w_1 - w^*\|_{\mathcal{H}}^2] \leq \frac{M^2}{\alpha^2}$$

since strong convexity implies

$$\begin{aligned} \langle w - w^*, \nabla f(w) \rangle &\geq \alpha \|w - w^*\|_{\mathcal{H}}^2 \\ \langle w - w^*, \nabla f(w) \rangle^2 &\geq \alpha^2 \|w - w^*\|_{\mathcal{H}}^4 \end{aligned}$$

and by Cauchy-Schwartz inequality we get

$$\|w - w^*\|_{\mathcal{H}}^2 \cdot \|\nabla f(w)\|_{\mathcal{H}}^2 \geq \langle w - w^*, \nabla f(w) \rangle^2$$

which yields

$$\begin{aligned} \|w - w^*\|_{\mathcal{H}}^2 \cdot \|\nabla f(w)\|_{\mathcal{H}}^2 &\geq \alpha^2 \|w - w^*\|_{\mathcal{H}}^4 \\ \|\nabla f(w)\|_{\mathcal{H}}^2 &\geq \alpha^2 \|w - w^*\|_{\mathcal{H}}^2 \end{aligned}$$

for all w . Taking exponents on both sides and the bound $\|\nabla f(w)\|_{\mathcal{H}}^2 \leq M^2$ we get

$$\mathbb{E}[\|w - w^*\|_{\mathcal{H}}^2] \leq \alpha^{-2}M^2$$

which concludes the proof using $\alpha = 1$ (Proposition 2). \square

Notice that $\|\nabla f(w)\|_{\mathcal{H}}^2 \leq M^2$ does indeed hold since $\nabla f(w) = w - \phi(X) \in H$. The following theorems describes the convergence rate of the objective function f in terms of the number of iterations t .

Theorem 2: Under the prerequisites of Theorem 1 it holds that

$$\mathbb{E}[f(w_t) - f(w^*)] \leq \frac{1}{2} \frac{M^2}{t}.$$

Proof: Using Eq. (2) and the bound for $Q(\theta)$ derived before yields the desired result. \square

We showed above that (in expectation) the distance between the optimal objective $f(w^*)$ and $f(w_t)$ decays as $\mathcal{O}(1/t)$. Another question is how this effects the EXPoSE decision rule $\eta(y) = \langle \phi(y), \mu[\mathbb{P}] \rangle$. By definition and the application of the Cauchy-Schwarz inequality

it holds that

$$\begin{aligned} \|\langle \phi(y), \mu[\mathbb{P}] \rangle - \langle \phi(y), w_t \rangle\| &= \|\langle \phi(y), \mu[\mathbb{P}] - w_t \rangle\| \\ &\leq \|\phi(y)\|_{\mathcal{H}} \cdot \|\mu[\mathbb{P}] - w_t\|_{\mathcal{H}} \end{aligned}$$

for all $y \in \mathcal{H}$. Taking expectations yields

$$\mathbb{E}[\|\langle \phi(y), \mu[\mathbb{P}] \rangle - \langle \phi(y), w_t \rangle\|^2] \leq \|\phi(y)\|_{\mathcal{H}}^2 \frac{M^2}{t}$$

for all $t \in \mathbb{N}$.

Algorithm 1 EXPoSE using Stochastic Optimization

Require:

1: T : the number of iterations *or* ϵ : accuracy

Algorithm:

2: Set $w_1 \leftarrow 0$

3: **for** $t \leftarrow 1, 2, \dots, T$ **do**

4: Sample x_t uniformly from \mathbb{P}

5: Set $\gamma_t \leftarrow \frac{1}{t}$

6: Set $\tilde{\nabla}f(w_t) \leftarrow w_t - \phi(x_t)$

7: Update $w_{t+1} \leftarrow w_t - \gamma_t \tilde{\nabla}f(w_t)$

8: Project $w_{t+1} \leftarrow w_{t+1} \cdot \max\{1, M\|w_{t+1}\|\}^{-1}$

9: **return** w_{T+1}

The stochastic optimization procedure for EXPoSE is summarized in Algorithm 1. Please note that the stochastic optimization procedure presented here is relatively simple and requires only a few lines of code to implement. It also does not introduce additional parameters since the optimal step-size is known. Step-sizes are crucial and difficult to determine in most optimization algorithms as they have a significant effect on the results. The bound M of the kernel is typically known and the number of iterations T determines the computing time and accuracy. Alternatively, the number of iterations T can be calculated given a desired accuracy ϵ using Theorem 1. The projection operator $\Pi_H(w)$ in the last step takes a form which can efficiently be computed, projecting w onto the sphere H .

We emphasize that the stochastic optimization procedure introduced here does not improve on the $\mathcal{O}(1/\sqrt{t})$ convergence rate of the empirical kernel mean map as demonstrated in Theorem 1, but introduces a methodology to reduce the computational complexity from linear to constant.

4.1 CONVERGENCE OF EXPOSE

Since $w \rightarrow w^*$ converges, this implies also the weak convergence [Pey15] from $w \rightarrow w^*$ namely

$$\lim_{t \rightarrow \infty} \langle u, w_t \rangle = \langle u, w^* \rangle, \quad \forall u \in H$$

and especially

$$\lim_{t \rightarrow \infty} \langle \phi(y), w_t \rangle = \langle \phi(y), \mu[\mathbb{P}] \rangle, \quad \forall y \in \mathcal{X}$$

which justifies the use of w_t as a surrogate for $\mu[\mathbb{P}]$.

4.2 REGULARIZATION

We would like to mention that the reformulation of EXPoSE as an optimization problem also introduces the opportunity to add constraints or similar properties to the objective function. One approach is to define a general *regularizer* $\lambda\Omega(w)$ on H replacing $\frac{1}{2}\langle w, w \rangle$ in Eq. (3) which yields

$$\min_{w \in H} \mathbb{E}[f(w)] = \min_{w \in H} \int \lambda\Omega(w) - \langle \phi(X), w \rangle \, d\mathbb{P}$$

with some regularization parameter $\lambda \geq 0$. An example would be to add a roughness penalty to the space of functions setting

$$\lambda\Omega(w) = \lambda \langle D^2 w, D^2 w \rangle$$

where D denote the differential operator. Another possibility is to place a sparsity constraint on w . If \mathcal{H} admits it, we can use

$$\lambda\Omega(w) = \lambda \|w\|_1,$$

where $\|\cdot\|_1$ is the l_1 -norm.

The disadvantage of other objective functions is, that these are in general not strongly-convex and hence yielding a slower convergence rate and may require additional parameters which are difficult to tune.

5 Experimental Evaluation

We present experimental results demonstrating the benefit of the proposed approach. Since the true distribution \mathbb{P} is often unknown and a closed form solution of $\mu[\mathbb{P}]$ is not available, we will use the empirical distribution \mathbb{P}_n as its surrogate in the objective function and measure the behavior of

$$\|w_t - \mu[\mathbb{P}_n]\|_{\mathcal{H}}$$

as t increases. For sufficiently large sample sizes n we can expect $\mu[\mathbb{P}_n]$ to be a good proxy for $\mu[\mathbb{P}]$ by the law of large numbers. Besides the convergence of the model $w_t \rightarrow \mu[\mathbb{P}]$, we will examine and compare the anomaly

detection scores

$$\eta_n(y) = \langle \phi(y), \mu[\mathbb{P}_n] \rangle \quad \text{and} \\ \eta_t(y) = \langle \phi(y), w_t \rangle$$

calculated by the empirical distribution (which is the original EXPoSE predictor proposed in [SEP15]) and the stochastic optimization approximation, respectively.

5.1 APPROXIMATE FEATURE MAPS

While it is theoretically possible to calculate quantities like $\|w_t - \mu[\mathbb{P}_n]\|_{\mathcal{H}}$ for any kernel k , this is extremely slow and intractable for most large-scale datasets. For datasets with a small sample size n we cannot expect $\mu[\mathbb{P}_n]$ to be a good proxy for $\mu[\mathbb{P}]$. We therefore omit an experiment with explicit features as we either cannot compute $\mu[\mathbb{P}_n]$ (large n) or $\mu[\mathbb{P}_n]$ is not a good estimate for $\mu[\mathbb{P}]$ (small n).

In order to overcome this problem, EXPoSE exploits the idea of *approximate feature maps* for its computational efficiency. The aim is to find approximations $\hat{\phi}: \mathcal{X} \rightarrow \mathbb{R}^r$ of ϕ such that

$$k(x, y) \approx \langle \hat{\phi}(x), \hat{\phi}(y) \rangle$$

for all $x, y \in \mathcal{X}$ and $r \in \mathbb{N}$. We will utilize the Random Kitchen Sinks (RKS) approach [RR07; RR08] which is based on Bochner's theorem for translation invariant kernels (such as the Gaussian RBF, Laplace, Matérn covariance, etc.). For example in the following experiments we will use the Gaussian RBF kernel $k(x, y) = \exp(-\frac{1}{2\sigma^2}\|x - y\|^2)$, which can be approximated by

$$Z \in \mathbb{R}^{r \times d} \text{ with } Z_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \hat{\phi}(x) = \frac{1}{\sqrt{r}} \exp(iZx),$$

where d is the dimension of $\mathcal{X} \subseteq \mathbb{R}^d$. The parameter $r \in \mathbb{N}$ determines the number of kernel expansions and is typically around 20,000. The specific choice of approximate feature map *does not* affect the previous theoretical analysis and other feature map approximations [LIS10; VZ12; KK12] can be used as well.

5.2 DATASETS

The following datasets, which all have purposely very different feature characteristics, are used to perform anomaly detection. We refer to [SEP15] for a detailed description of the datasets and feature characteristic.

- The MNIST database contains 70,000 images of

handwritten digits. Using the raw pixel values yield an input space dimension of 784.

- KDD-CUP 99 is an intrusion detection dataset which contains 4,898,431 connection records of network traffic. As in [SEP15] we rescale the 34 continuous features to $[0, 1]$ and apply a binary encoding for the 7 symbolic features.
- The third dataset contains 600,000 instances of the *Google Street View House Numbers* (SVHN) [Net+11] where we use the *Histogram of Oriented Gradients* (HOG) with a cell size of 3 to get a 2592-dimensional feature vector.

The kernel bandwidth σ^2 used for these datasets are 7.0, 5.6 and 7.8 respectively, which we found to yield a reasonable anomaly detection performance.

Since SVHN and MNIST are multi-class and not anomaly detection datasets we use the digit 1 as *normal* class and all other digits as *anomaly* instances.¹ At each iteration of Algorithm 1 we uniformly choose an instance from the (training) dataset not used previously. We then update the model w_t according to the algorithm. Every 200 iterations, w_t is used to calculate an anomaly detection score for 10,000 dedicated random instances of the (test) dataset using the full model $\eta_n(y)$ and the stochastic optimization approximation $\eta_t(y)$.

5.3 DISCUSSION

The experimental results with approximate feature maps are shown in Fig. 2. The first row contains traces of the objective function $f(w_t) - f(w^*)$, where $w^* \approx \mu[\mathbb{P}_n]$ for all three datasets. The stochastic optimization algorithm already reaches a reasonable low objective after a few hundred iterations. A further improvement is only visible on a logarithmic scale (dashed blue) on the second y-axis on the right. More important, we observe a similar effect in the second row when comparing $\|w_t - w^*\|$. We get near to w^* relatively fast, but it takes much more samples to estimate w^* with a high accuracy. However, we will see that a high accuracy estimation is necessary for a good anomaly detection performance. To measure the anomaly detection rate, we first plug w_t and w^* into the EXPoSE estimators $\eta_t(y)$ and $\eta_n(y)$ respectively and calculate scores for all instances in the test dataset. The difference of these scores are shown in row number three. We see again, that the stochastic optimization approximation $\eta_t(y)$ yields similar scores as the full $\eta_n(y)$. The last row illustrates the development of the classification error as

more iterations are performed². After only a few hundred iterations $\eta_t(y)$ reaches the same classification error as the original EXPoSE predictor $\eta_n(y)$. This confirms that a high accuracy approximation of w^* does not necessarily lead to a better predictor. The key is, that for a given ϵ we can reach $\|w_t - w^*\| < \epsilon$ in a fixed number of iterations, *independent* of the dataset size n which reduced the computational complexity from $\mathcal{O}(n)$ to $\mathcal{O}(1)$.

We emphasize that, unlike other regularized risk minimization problems, EXPoSE does not have a regularization parameter. This is important as the authors of Pegasos noticed that “[...] the runtime to achieve a predetermined suboptimality threshold would increase in proportion to λ [the regularization parameter]. Very small values of λ (small amounts of regularization) result in rather long runtimes” [SS+11].

6 Conclusion

In this work we cast the EXPoSE anomaly detection algorithm into a stochastic optimization problem. This enables us to fine an ϵ -accurate approximation of the kernel mean map $\mu[\mathbb{P}]$ in *constant time*, independent of the training dataset size n . In particular, this approximations reduces the computational complexity of EXPoSE and the empirical kernel mean map from the previous $\mathcal{O}(n)$ to $\mathcal{O}(1)$ whenever an ϵ -accurate estimation is sufficient. More precisely, we are able to determine the number of necessary stochastic optimization iterations T for a user defined error threshold ϵ such that $\|w_T - w^*\| < \epsilon$. The intuition is that a very high accuracy estimation w^* does not necessarily result in a better anomaly detection performance and hence there is no benefit in spending more computational resources. This intuition is also confirmed experimentally on three large-scale datasets, where we reach the same anomaly detection performance long before all data is incorporated into the model. This is the first time that an optimization routine is used for EXPoSE and we provide a detailed theoretical analysis of this algorithm. We emphasize that the proposed algorithm does not introduce any additional parameters which have to be tuned and the gradient descent step-sizes are determined automatically. This has significant implications for large-scale applications such as anomaly detection problems and other techniques which are based on the kernel mean embedding.

²The prediction score threshold is determined by means of cross-validation.

¹A different normal/anomaly setup had no significant impact on the experimental results.

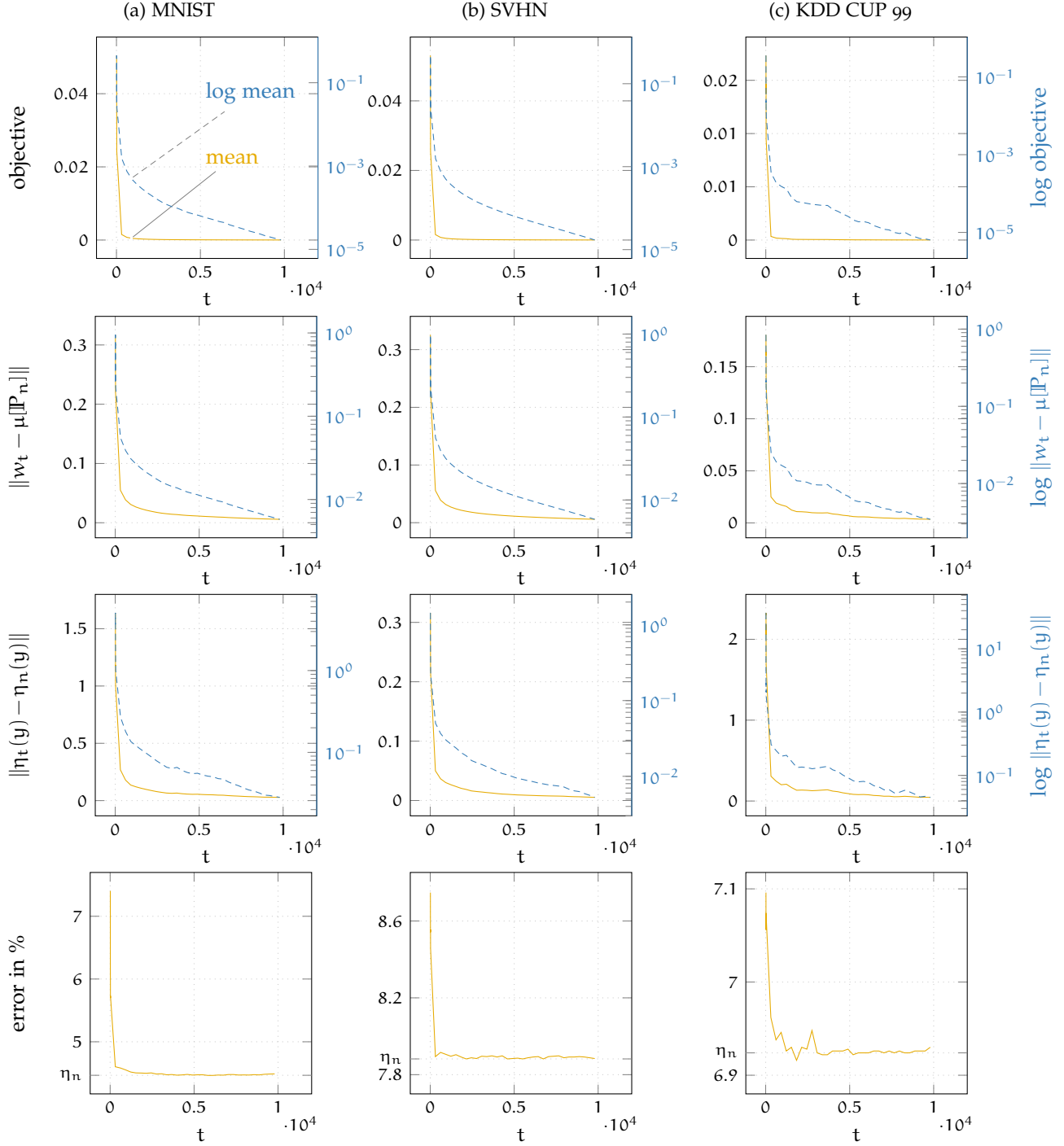


Figure 2: Evaluation of the stochastic optimization approach for EXPoSE. The datasets are organized in columns. The first row illustrates the difference between objective functions $f(w_t) - f(w^*)$. In the second row we show the deviation of w_t from $\mu[\mathbb{P}_n]$ as $\|w_t - \mu[\mathbb{P}_n]\|$. In the third row we plotted the difference in scores $\|\eta_t(y) - \eta_n(y)\|$, averaged over all query points y in the test dataset. The last row shows the anomaly detection performance of EXPoSE. In all figures, the solid line is the mean over 10 experiments and on the second y-axis on the right we show the same curve (dashed) on a logarithmic scale when appropriate.

References

- [Aga+11] A. Agarwal et al. "Information-Theoretic Lower Bounds on the Oracle Complexity of Convex Optimization Convex optimization". In: *Advances in Neural Information Processing Systems*. 2011, pp. 1–9.
- [Bac14] F. Bach. "Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 595–627.
- [BB08] O. Bousquet and L. Bottou. "The trade-offs of large scale learning". In: *Advances in neural information processing systems*. 2008, pp. 161–168.
- [BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [CBK09] V. Chandola, A. Banerjee, and V. Kumar. "Anomaly detection: A survey". In: *ACM Computing Surveys (CSUR)* 41.3 (2009), pp. 1–58.
- [Gre+05] A. Gretton et al. "Measuring statistical dependence with Hilbert-Schmidt norms". In: *Algorithmic learning theory*. Springer. 2005, pp. 63–77.
- [Gre+12] A. Gretton et al. "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.
- [JN10] A. Juditsky and Y. Nesterov. "Primal-dual subgradient methods for minimizing uniformly convex functions". In: *Universite Joseph Fourier, Tech. Rep* (2010).
- [KK12] P. Kar and H. Karnick. "Random feature maps for dot product kernels". In: *International Conference on Artificial Intelligence and Statistics* (2012), pp. 583–591.
- [LIS10] F. Li, C. Ionescu, and C. Sminchisescu. "Random Fourier approximations for skewed multiplicative histogram kernels". In: *Pattern Recognition*. Springer, 2010, pp. 262–271.
- [Nem+09] A. Nemirovski et al. "Robust stochastic approximation approach to stochastic programming". In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609.
- [Nes04] Y. Nesterov. *Introductory lectures on convex optimization*. Vol. 87. Springer Science & Business Media, 2004.
- [Nes83] Y. Nesterov. "A method of solving a convex programming problem with convergence rate $O(1/k_2)$ ". In: *Soviet Mathematics Doklady* 27.2 (1983), pp. 372–376.
- [Net+11] Y. Netzer et al. "Reading digits in natural images with unsupervised feature learning". In: *NIPS workshop on deep learning and unsupervised feature learning*. Vol. 2011. 2011, p. 4.
- [Pey15] J. Peypouquet. *Convex Optimization in Normed Spaces: Theory, Methods and Examples*. Springer, 2015.
- [RM51] H. Robbins and S. Monroe. "A stochastic approximation method". In: *The annals of mathematical statistics* (1951), pp. 400–407.
- [RR07] A. Rahimi and B. Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems*. 2007, pp. 1177–1184.
- [RR08] A. Rahimi and B. Recht. "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". In: *Advances in neural information processing systems*. 2008, pp. 1313–1320.
- [RSB12] N. L. Roux, M. Schmidt, and F. R. Bach. "A stochastic gradient method with an exponential convergence rate for finite training sets". In: *Advances in Neural Information Processing Systems*. 2012, pp. 2663–2671.
- [SEP15] M. Schneider, W. Ertel, and G. Palm. "Expected Similarity Estimation for Large Scale Anomaly Detection". In: *International Joint Conference on Neural Networks*. IEEE, 2015, pp. 1–8.
- [SFL11] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. "Universality, characteristic kernels and RKHS embedding of measures". In: *The Journal of Machine Learning Research* 12 (2011), pp. 2389–2410.
- [Smo+07] A. J. Smola et al. "A Hilbert space embedding for distributions". In: *Algorithmic Learning Theory*. Springer. 2007, pp. 13–31.
- [SS+11] S. Shalev-Shwartz et al. "Pegasos: Primal estimated sub-gradient solver for SVM". In: *Mathematical Programming*. Vol. 127. 1. 2011, pp. 3–30.
- [SST11] S. Shalev-Shwartz and A. Tewari. "Stochastic methods for l_1 -regularized loss minimization". In: *The Journal of Machine Learning Research* 12 (2011), pp. 1865–1892.
- [VZ12] A. Vedaldi and A. Zisserman. "Efficient Additive Kernels via Explicit Feature Maps". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34.3 (2012), pp. 480–492.